ELSEVIER

Nuclear Instruments and Methods in Physics Research A 480 (2002) 713–725

NUCLEAR INSTRUMENTS & METHODS IN PHYSICS RESEARCH Section A

www.elsevier.com/locate/nima

Prospects for charge sensitive amplifiers in scaled CMOS ☆

Paul O'Connor*, Gianluigi De Geronimo

Brookhaven National Laboratory, Instrumentation Division, Bldg 535B, Upton, NY 11973, USA

Received 6 February 2001; received in revised form 26 March 2001; accepted 27 March 2001

Abstract

Due to its low cost and flexibility for custom design, monolithic CMOS technology is being increasingly employed in charge preamplifiers across a broad range of applications, including both scientific research and commercial products. The associated detectors have capacitances ranging from a few tens of fF to several hundred pF. Applications call for pulse shaping from tens of ns to tens of μs, and constrain the available power per channel from tens of μW to tens of mW. At the same time a new technology generation, with changed device parameters, appears every 2 years or so. The optimum design of the front-end circuitry is examined taking into account submicron device characteristics, weak inversion operation, the reset system, and power supply scaling. Experimental results from recent prototypes will be presented. We will also discuss the evolution of preamplifier topologies and anticipated performance limits as CMOS technology scales down to the 0.1 μm/1.0 V generation in 2006.

PACS: 07.50.E; 84.30; 85.40

Keywords: Charge sensitive amplifier; CMOS scaling

1. Introduction

Charge sensitive amplifiers (CSAs) are used extensively in processing the signals from capacitive sensors such as photodetectors, pressure sensors, particle and X-ray detectors, and pyroelectric detectors. Increasingly, these amplifiers are being implemented in monolithic processes where there is a need for high-volume production (e.g. particle physics collider detectors), for interfacing to a dense array of sensors (pixel detectors), or whenever miniaturization and high functional integration are at a premium. The high input impedance of the MOSFET makes CMOS an attractive technology for fabricating such amplifiers, particularly if a high level of integration is desired. However, CMOS technology development is driven by the needs of digital VLSI, and the resulting rapid feature size scaling presents several challenges for high dynamic range CSAs:

- increase in MOSFET noise due to carrier heating in the channel, higher interface trap density, gate tunneling current, and larger parasitic resistance;
- reduced power supply voltage which restricts the output swing, constrains the circuit topology, and increases the noise of current sources;

☆Work supported by the US Department of Energy, Contract No. DE-AC02-98CH10886.

*Corresponding author. Tel.: +1-631-344-7577; fax: +1-631-344-5773.

E-mail address: poc@bnl.gov (P. O'Connor).

0168-9002/02/$ - see front matter
PII: S0168-9002(01)01212-8

- increased application demand for mixed-signal circuits having digital switching activity occurring on the same substrate with highly sensitive CSAs;
- decreased availability of quality passive components for analog design;
- poor modeling of the DC, AC, and noise properties of the devices.

Many of these drawbacks have been discussed in the overall context of analog design in scaled CMOS [1,2]. In the following sections, we consider only those aspects of CMOS scaling which impact the performance of CSAs. After a review of CSA operation and CMOS scaling in Sections 2 and 3, Section 4 covers the noise sources in MOSFETs and Section 5 examines the expected scaling of the input-referred noise charge and dynamic range. Section 6 discusses sources of noise other than the input transistor and additional scaling effects.

2. CSA operation

Charge measurement systems are characterized by system requirements which vary tremendously from application to application:

- dynamic range 6–20 bits;
- sensor capacitance 50 fF–10 nF;
- speed of response 5 ns–1 ms;
- power dissipation 10 μW–100 mW.

A generic block diagram, shown in Fig. 1, represents such a system. The sensor, with capacitance C_{det}, produces pulses of charge that are integrated on the feedback capacitor C_f. The amplifier is characterized by a series input voltage noise source e_n with a white component set by the input device transconductance g_m and a $1/f$ component that is inversely proportional to the device area WL.

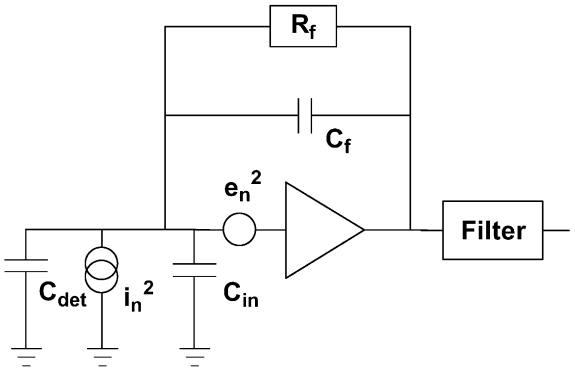

Fig. 1. Simplified CSA schematic.

In the case of CMOS the equivalent input noise current is negligible; however the secondary feedback element R_f which discharges the feedback capacitor contributes noise whose impact must be carefully considered. A filter amplifier following the preamplifier is necessary to eliminate out-of-band noise. The resolution of the system is expressed by the equivalent input noise charge (ENC) [3,4]:

$$\mathrm{ENC}^2 = (C_{det} + C_{in})^2 \left(\frac{a_1 2kTR_s}{t_s} + \frac{a_2 \pi K_f}{C_{ox} WL} \right) + \frac{a_3 2kTt_s}{R_p} \quad (1)$$

where C_{in} is the capacitance of the input transistor, R_s is its equivalent series noise resistance, R_p is the effective noise resistance of the feedback element R_f, t_s is the characteristic time constant of the post-filter (shaping time), K_f is the $1/f$ noise coefficient of the input transistor, and a_1, a_2, and a_3 are form factors (typically close to 1) related to the series white, series $1/f$, and parallel white noise respectively. Typically series white noise dominates for short shaping times, R_p for long shaping times, and $1/f$ noise in the intermediate range. High capacitance detectors exacerbate the series noise terms. While both series white noise and parallel noise can be combated by circuit techniques, expenditure of more power, etc., the effect of $1/f$ noise on CSA performance is fundamentally limiting. In CMOS, the lowest-noise CSAs are always dominated by the $1/f$ noise properties of the input transistor.

An ideal CSA technology would have a high g_m/C_{gs} ratio (f_T) at low current, as well as low γ ($\gamma = R_s \times g_m$) for minimizing the white series noise; a low $1/f$ noise coefficient K_f; and controllable sub-nA current sources for low parallel noise. A high quality floating capacitor is needed for the preamp integrating capacitor and

for the post-filter. Other desirable features for CSAs are:

- high g_m/g_d for amplifier gain;
- excellent AC isolation for integration with digital circuits;
- high supply voltage for driving subsequent stages and for cascodes;
- ESD—tolerant;
- radiation tolerant.

In contrast, the economic realities of semiconductor fabrication compel analog designers to use commercially available processes that depart only minimally from standard digital VLSI production. Such processes are presently in a state of rapid evolution towards higher integration density.

3. Digital CMOS scaling

CMOS technology scaling takes place by a series of well-defined process generations, coordinated among foundries and equipment manufacturers. In each process generation, integration density doubles and speed increases by 50%. Constant-voltage scaling, which had been followed up to the 0.8 μm generation, has given way to quasi-constant field scaling as oxide and junction breakdown limits have been reached. In constant field scaling all dimensions are multiplied by the scale factor λ, doping density is increased, and supply voltage scaled down by the same factor. Submicron device scaling below 0.5 μm poses several major challenges to MOS device engineering, the most important of which are to minimize undesirable short-channel effects, to control power dissipation, and to ensure reliability to the level of one failure in 10^7 chip-hours of operation.

The roadmap for the next several technology generations is shown in Table 1 [5].

4. Device noise trends

4.1. Series white noise

For long channel MOSFETs the proportionality constant γ relating noise to transconductance ($R_s = \gamma/g_m$) has the value 2/3 in strong inversion, 1 in the linear region, and 1/2 in weak inversion [6,7] (here the contribution from the bulk transconductance g_{mb} has been neglected). In short channel devices the carriers can acquire enough energy from the electric field in the channel to raise their effective temperature above that of the lattice. Models of this heating effect [8,9] lead to values of γ greater than those for long channel devices. However, the predictions of different models are inconsistent, or are only supported by experimental data over a limited range of bias conditions. Predictions are particularly lacking for the normal bias point for a device used as the input transistor of a CSA, namely low drain-source voltage (just above V_{DS-sat}) and low current density I_D/W (moderate inversion). Values of γ as high as 2–4 have been reported [10,11], but these results are from experimental short channel devices that do not follow the same scaling laws as current commercial devices (e.g., L_{min}/t_{ox} ratio). Also high γ values are usually reported on devices

Table 1
CMOS technology roadmap

Year	1997	1999	2001	2003	2006	2009
Feature size (μm)	0.25	0.18	0.15	0.13	0.10	0.07
Supply (V)	2.5	1.8	1.6	1.5	1.2	0.9
Tox (nm)	5.0	4.0	3.3	2.8	2.2	2.0
Vth (mV)	500	470	440	420	400	370
Nsub (10^{16}/cm^3)	3.4	5	6	7	10	20
Xj (nm)	100	70	50	<50	<50	<50
10^6 FET/cm^2	8	14	16	24	40	64
Interconnect (km/chip)	0.82	1.5	2.2	2.8	5.1	10

at high V_{DS} and high I_D/W. Hence, we have measured noise on MOS transistors fabricated in representative commercial submicron processes, for devices in a typical CSA bias condition. These results (along with recently reported results from similar devices [12–14]) are shown in Fig. 2, for four technologies with a minimum gate length L_{min} from 0.7 to 0.25 μm. It can be seen that for devices from the same process there is a modest increase in γ as the gate length decreases. However, there is no trend towards higher γ for the smallest-L_{min} devices in each technology.

Another source of series thermal noise are the parasitic source and drain resistance (R_S, R_D) of the MOSFET. Scaled CMOS requires shallow junctions whose high resistance has been seen as a possible cause of increased noise. But unrestricted growth of R_S and R_D would be detrimental to digital performance as well. It can be expected that the parasitic resistance will be held to values much less than $1/g_m$ through silicidation and heavy doping of the contacts in the interest of maintaining logic speed; this will assure a minimal contribution to white series noise.

4.2. Series 1/f noise

High and variable $1/f$ noise has always been characteristic of MOSFETs. It is known to be strongly dependent on interface quality and gate processing. In long channel FETs, PMOS devices have 3–30 times lower $1/f$ noise power spectrum than equally sized NMOS for reasons that may be related to buried channel conduction.

For deep submicron processes, the PMOS will be formed using p+ poly gates and retrograde well doping, causing the inversion layer centroid to be located closer to the Si–SiO_2 interface. The change of the PMOS from a buried channel to a surface channel device is predicted to lead to an increase in K_f to a value near that of NMOS. The relative advantage of using PMOS in $1/f$ noise-sensitive applications would then disappear. Although evidence for this effect has been reported for 0.25 and 0.18 μm devices measured at very low frequencies [2,15], this is not confirmed by our recent measurements on devices down to the 0.25 μm generation. As shown in Fig. 3, the scaled devices exhibit no significant increase in K_f for either NMOS or PMOS. The PMOS retains its $\approx 10\times$ advantage over NMOS in these particular processes.

The shallow junctions required for scaled devices can only be preserved by limiting the thermal budget—hence gate processes in deep submicron devices will have reduced post-oxidation anneal and higher trap density. For ultrathin gate dielectrics, new materials with higher trap densities than SiO_2 will be used (nitrided, halogenated, H_2 annealed). These alternative methods of gate oxide formation have been shown to increase

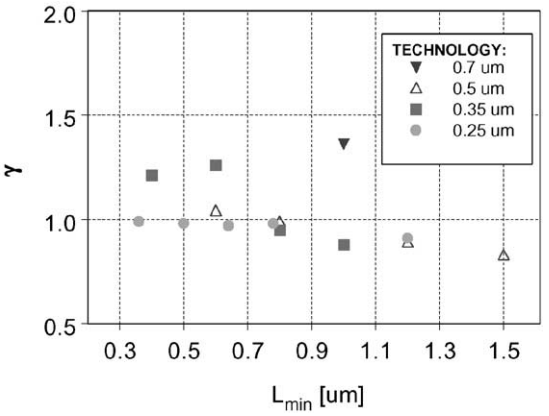

Fig. 2. Thermal noise coefficient γ for NMOS devices from several commercial submicron processes.

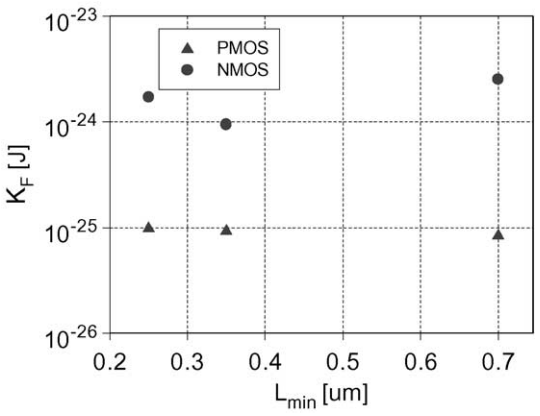

Fig. 3. $1/f$ noise coefficient K_f vs. minimum feature size for NMOS and PMOS transistors from 3 submicron foundries.

the $1/f$ noise by more than one order of magnitude [15].

Hot carrier stress is another mechanism that introduces noise-producing trap states in submicron MOSFETs [15,16]. The scaled MOSFET is engineered to have acceptable degradation of DC characteristics (g_m, V_{th}) over the expected lifetime. However, $1/f$ noise is found to increase far more rapidly with hot carrier stress than the DC parameters. For example, a device in which g_m degraded 10% during a 7-h over-voltage stress exhibited a 400% increase in low frequency noise [16]. The effect is found to be worse for short channel devices and strongly dependent on bias condition during stress. Submicron MOSFETs experience some level of hot carrier stress during normal operation at the permitted supply voltage. Hence there is a possibility that devices that have been engineered for "acceptable" degradation of DC parameters may experience an unacceptable increase in $1/f$ noise.

4.3. Gate current

At a bias of 1.5 V, gate current density increases by 10 orders of magnitude as the oxide thickness decreases from 3.6 to 1.5 nm [17]. This corresponds to the expected oxide thickness change from 0.15 to 0.07 μm generation shown in Table 1. A gate current density of $I_g = 1\,A/cm^2$ is considered tolerable for digital circuits based on power dissipation considerations (total gate area per chip $\sim 0.1\,cm^2$). At this current density, a typical CSA input FET optimized for low series noise would have I_g of the order 10–100 nA. A parallel noise contribution to the ENC of 200–700 rms electrons would result for system with 1 μs shaping time. To optimize noise, the selection of the device geometry (see Section 5) would have to consider the simultaneous minimization of series and parallel noise.

5. ENC and dynamic range scaling

For many applications, series white noise is the dominant source. Its scaling properties are influenced not only by the coefficient γ but also by the device cutoff frequency f_T and supply voltage. To obtain the lowest ENC it is first necessary to determine the optimum size of the input transistor.

5.1. Dimensioning the input device: generalized capacitive match condition

In Eq. (1), the series white component of the ENC is given by

$$ENC_{sw}^2 = (C_{det} + C_{ox}WL)^2 \frac{a_1 2kT\gamma}{g_m t_s} \tag{2}$$

where $C_{in} = C_g = C_{ox}\ WL$. We seek the device dimensions W and L that will minimize Eq. (2), under the constraint of constant drain current I_D.

Most charge preamplifiers are designed to operate with the input transistor in moderate inversion and at low V_{DS}, to maximize the g_m/I_D ratio and dynamic range. Under these operating conditions, excess noise at short channel lengths has been found to be a minor effect (see Fig. 2). In these technologies, the optimum channel length will always be L_{min}, since this value maximizes the transconductance to capacitance ratio of the device.

It is well known that the series white noise has a minimum with respect to the channel width, since increasing W causes the input equivalent voltage noise generator ($\sim 1/g_m$) to decrease, while simultaneously increasing the total input capacitance. To find the optimum W, we must recognize that the dependence of g_m on W changes depending on whether the device is operated in weak inversion, strong inversion, or velocity saturation. The corresponding relations for g_m are [21]

$$g_m = \frac{I_D}{nV_t}, \quad \left(\frac{I_D}{W}\right) < \frac{2\mu C_{ox}(nV_t)^2}{L} \tag{3a}$$

$$g_m = \sqrt{2\mu C_{ox}\frac{W}{L}I_D}, \quad \frac{2\mu C_{ox}(nV_t)^2}{L} < \left(\frac{I_D}{W}\right) < \frac{C_{ox}Lv_{sat}^2}{2\mu} \tag{3b}$$

$$g_m = C_{ox}Wv_{sat}, \quad \left(\frac{I_D}{W}\right) > \frac{C_{ox}Lv_{sat}^2}{2\mu} \tag{3c}$$

where I_D is the drain current, μ is the carrier mobility, W and L are the channel width and length respectively, v_{sat} is the carrier saturation velocity, n is the subthreshold slope factor and

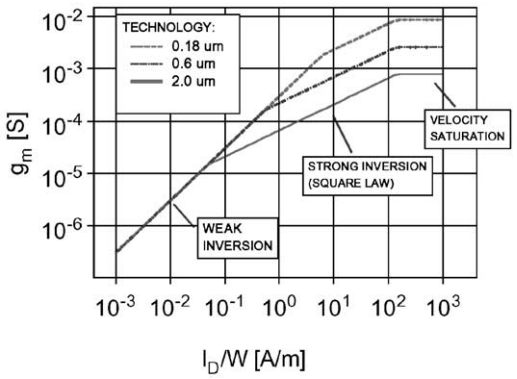

Fig. 4. NMOS transconductance as a function of current density for several CMOS generations.

$V_t = kT/q$. Fig. 4 illustrates the dependence of NMOS g_m on drain current density, showing three regions in which

$$g_m \sim \left(\frac{I_D}{W}\right)^{\alpha} \tag{4}$$

where $\alpha = 1$, 1/2, and 0 in the weak inversion, strong inversion (square-law) and velocity saturated regions, respectively. The strong inversion region is usually considered the normal operating condition for an analog MOSFET; however, under continued scaling, the weak inversion regime will extend to higher and higher current densities. By the 0.13 μm generation, the weak inversion regime will extend directly into the velocity saturation region, and the square-law regime will vanish altogether.

Optimum capacitive matching of MOSFET-input CSAs to detectors has been considered by Chang and Sansen [18,19] and by Bertuccio [20]. Their analyses consider only the case where the MOSFET is in the strong inversion (square-law) region of operation, in which case the minimum noise occurs when the FET input capacitance equals one third of the detector capacitance. A generalization of the optimum matching for arbitrary operating point is given in Appendix A, and leads to the results

$$\text{Region I: } \frac{C_{det}}{I_D} < \frac{6\mu}{v_{sat}^2} \text{ (velocity saturated)}$$
$$C_{g,opt} = C_{det}. \tag{5}$$
$$\text{Region II: } \frac{6\mu}{v_{sat}^2} < \frac{C_{det}}{I_D} < \frac{3L_{min}^2}{2\mu(nV_t)^2} \text{ (strong inversion)}$$
$$C_{g,opt} = \frac{C_{det}}{3}. \tag{6}$$

$$\text{Region III: } \frac{C_{det}}{I_D} > \frac{3L_{min}^2}{2\mu(nV_t)^2} \text{ (weak-strong inversion boundary)}$$
$$C_{g,opt} = \frac{L_{min}^2 I_D}{2\mu(nV_t)^2}. \tag{7}$$

In the above expressions L_{min} is the minimum gate length.

The following important conclusions stem from Eqs. (5)–(7):

- there are three separate noise matching relations, corresponding to the three MOSFET operating regions;
- the optimum matching depends only on the ratio of detector capacitance to available drain current; and
- weak inversion is always sub-optimal for MOSFET CSAs.

For high current densities the device is velocity-saturated and the cutoff frequency becomes independent of size; hence the optimum gate width is the one which gives $C_{g,opt} = C_{det}$. For moderate current density the input FET operates in the strong inversion (square-law) region and the optimum gate width is the one which gives $C_{g,opt} = C_{det}/3$, as derived in previous work [18–20]. For lower current densities the device should be sized so that it is operating at the boundary of weak and strong inversion. In weak inversion, the transconductance is independent of device geometry, hence the $g_m/(C_g + C_{det})$ ratio can always be improved by decreasing the device width. In Appendix A we show that the noise optimum occurs at the weak inversion/strong inversion boundary for such devices. The optimum input device capacitance, expressed as a fraction of the detector capacitance, is illustrated in Fig. 5 for a wide range of the parameter C_{det}/I_D.

5.2. ENC_s scaling

Now let us consider the evolution of the ENC of charge amplifiers fabricated in scaled CMOS technology. In Appendix B, we derive the following relation for the white series noise of a CSA

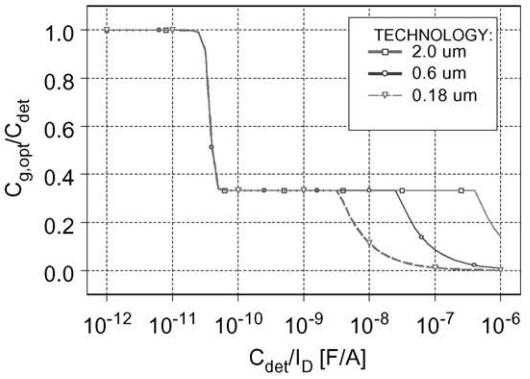

Fig. 5. Optimum input capacitance for white series noise as a function of the parameter C_{det}/I_D for NMOS devices in three different technology generations.

operated under the constraint of constant power dissipation

$$\mathrm{ENC}^2_{\mathrm{sw,min}} = \xi kTC_{\mathrm{det}}\frac{1}{t_s}L_{\mathrm{min}}\sqrt{\frac{C_{\mathrm{det}}V_{\mathrm{DD}}}{\mu P}} \tag{8}$$

where ξ is a numerical constant, t_s is the shaping time, and $P = V_{DD}I_D$ is the power dissipation of the input branch. The parameters that scale in the above expressions are

$$L'_{\mathrm{min}} \to \lambda L_{\mathrm{min}} \quad \text{and} \quad V'_{\mathrm{DD}} \to \lambda V_{\mathrm{DD}} \tag{9}$$

where λ is the feature size scaling factor. In Regions II and III therefore

$$\mathrm{ENC}'_{\mathrm{sw,min}} \to \lambda^{3/4}\mathrm{ENC}_{\mathrm{s,min}} \tag{10}$$

This means that for a given system, we can expect the series noise to improve by 23% for every CMOS generation (the historical scaling factor is $\lambda = 0.7$).

Alternatively, the results of Subsection A can be solved for the power required to achieve a given ENC. In this case we get

$$P' \to \lambda^3 P \tag{11}$$

and so a 60% decrease in power is expected per generation.

For devices operating in Region I the minimum ENC_s is given by

$$\mathrm{ENC}^2_{\mathrm{sw,min}} = \xi' kTC_{\mathrm{det}}\frac{1}{t_s}\frac{L_{\mathrm{min}}}{v_{\mathrm{sat}}} \tag{12}$$

which scales as

$$\mathrm{ENC}'_{\mathrm{sw,min}} \to \lambda^{1/2}\mathrm{ENC}_{\mathrm{s,min}} \tag{13}$$

or 16% per generation. In the velocity saturated case the power required to achieve a given ENC is independent of λ.

5.3. Optimization of total noise

Thus far we have considered only the white series noise. However, in MOS design $1/f$ noise is not negligible and has an important effect on capacitive matching and ENC scaling. As shown in Ref. [18–20], $1/f$ noise is minimized for a device whose input capacitance is equal to that of the detector. The total device noise always includes contributions from both white and $1/f$ sources. To minimize this total noise we arrive at an optimum gate capacitance which lies somewhere between C_{det} and one of the values given in Eqs. (5)–(7). The relative weight of the white and $1/f$ terms depends on shaping time and available power.

To make the scaling principles more concrete, let us consider four specific cases corresponding to common detector-amplifier combinations. For each case, system parameters capacitance (C_{det}), shaping time (t_s), power (P_{diss}), and detector leakage (I_{leak}) will be defined—see Table 2. Then the input transistor design will be optimized (choice of NMOS/PMOS, width) taking into account the white and $1/f$ noise sources using a simple mathematical model. This procedure will be repeated for the CMOS generations from 2 μm down to 0.1 μm. At each generation, the minimum gate length is selected. We assume that the noise coefficients γ and K_f will remain unchanged from the values commonly reported for today's technology, i.e. $\gamma \sim 0{\cdot}7$, $K_f \sim 10^{-24}$ J (NMOS), $K_f \sim 10^{-25}$ J (PMOS).

Results of this simulation are shown in Fig. 6, giving the minimum ENC, and Fig. 7, giving the optimum device size ($C_{g,opt}/C_{det}$) as functions of the minimum channel length L_{min}. In both figures, optimum device gender is indicated by the symbol (circle = NMOS, triangle = PMOS). These

Table 2
System parameters for ENC scaling simulation

System	C_{det}	t_s	P	I_{leak}	Detector	Typical application
a	30	75	10	0.001	Wire chamber	Tracking, imaging
b	15	25	0.2	10	Si Strip	Tracking
c	0.3	25	0.02	1	Si Pixel	Tracking
d	3	2500–500[a]	10	0.01	Semiconductor	Spectroscopy
UNITS	pF	ns	mW	nA	—	—

[a] For this system, the shaping time was varied at each L_g to optimize the overall noise (i.e., to make the white series noise and parallel noise equal).

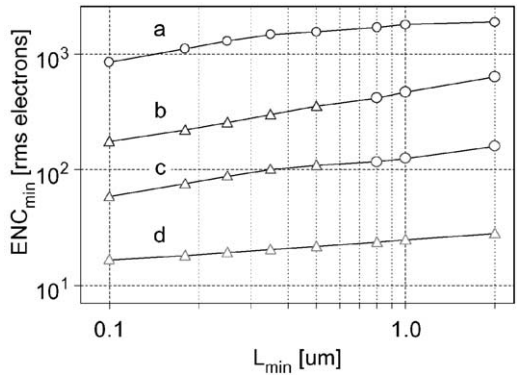

Fig. 6. Optimum total ENC as a function of the minimum gate length. Curves (a–d) refer to the system parameters given in Table 2. Circles = NMOS, triangles = PMOS.

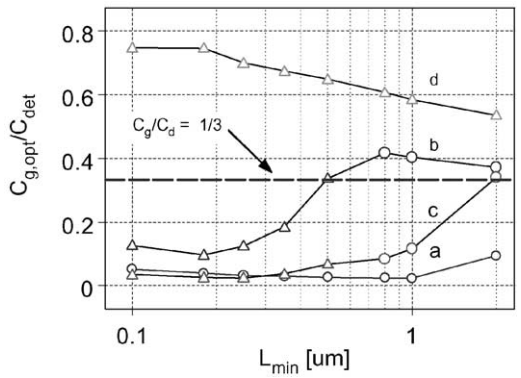

Fig. 7. The ratio $C_{g,opt}/C_{det}$ for optimum total ENC. Curves (a–d) refer to the system parameters given in Table 2. Circles = NMOS, triangles = PMOS.

examples illustrate the variety of effects contributing to scaled CMOS CSA optimization. Tradeoff of white vs. $1/f$ noise, switchover into weak inversion $g_m - I_D$ characteristics, and the wide range of system parameters C_{det}, t_s, and P have an effect on device sizing. Fig. 7 shows that the ratio $C_{g,opt}/C_{det}$ can vary from 1% to 75% depending on system parameters and scaling. In one case (Fig. 7c) device scaling alone shifts the optimum match point from 34% to 4%, almost an order of magnitude change. All the systems show the effects of a mixture of white and $1/f$ noise. The $1/f$ noise of a given gender of device is almost unchanged by scaling and is the absolute lower limit on noise in any MOS system [18,19]. The relative importance of $1/f$ noise is increased for long t_s, high P, and small L_{min}.

White series noise improves with scaling for two reasons: at shorter channel lengths the FET's cutoff frequency improves, and scaled technologies use reduced supply voltages, hence the current in the input branch can be increased without increasing power dissipation. This increased current further improves the device noise. Systems where white noise is most dominant (systems (a–c)) in Fig. 7 with $L_{min} \leqslant 0.4\,\mu m$) have ENC scaling behavior close to the $L_{min}^{3/4}$ dependence predicted by Eq. (10).

The improvement of white series noise at small L_{min} tends to increase the relative importance of the underlying $1/f$ noise. This has two effects. First, in some systems it becomes possible to switch the input device from NMOS to PMOS as L_{min} goes down (systems (a–c)). The high white noise of PMOS, which was prohibitive in the older generations, becomes more manageable with scaling and then its lower $1/f$ noise gives PMOS the advantage. A second effect is that the optimum C_g/C_{det} will go up as $1/f$ noise becomes higher weighted. This effect is seen most clearly in Fig. 7, system (d), although it also contributes in systems (a–c).

The switchover from strong to weak inversion operation at smaller L_{min} tends to push the ratio $C_{g,opt}/C_{det}$ below the classical square-law value of 1/3. Very small values of $C_{g,opt}/C_{det}$ can be seen in systems (b) and (c) where the C_{det}/I_D ratio is highest.

5.4. Dynamic range scaling

To calculate the maximum signal that can be processed, consider that the maximum output signal is $<V_{DD}$ and thus the maximum input charge is

$$Q_{in\ max} = cC_{det}V_{DD} \tag{14}$$

where $c = C_f/C_{det}$ is the inverse of the amplifier's charge gain. In practical circuits $c<0.1$. Now the maximum signal to noise ratio can be expressed, for the case of strong inversion, as

$$SNR = \frac{c(\mu PC_{det}V_{DD})^{1/4}}{\sqrt{10a_1kTL_{min}^{1/2}/t_s}}. \tag{15}$$

Substituting scaled quantities for V_{DD}, L_{min} we find that

$$SNR' \rightarrow \lambda^{1/4}SNR. \tag{16}$$

The SNR degrades 10% per generation since the decrease in supply voltage offsets the improvement in ENC found in the previous Section.

To compensate for this decline of SNR, the following solutions should be adopted:

- further reduce the ENC by increasing the power in the front device;
- replace conventional amplifier stages with rail-to-rail stages, especially in the output stage;
- convert single ended signals to differential, recovering a factor of two in signal swing. Note that this is usually accompanied by an increase in power and noise.

The first option above requires power to scale as

$$P' \rightarrow \lambda^{-1}P \tag{17}$$

or 43% increase per generation.

6. Other effects

6.1. Off-state leakage current

The MOSFET off-state leakage current I_{off} shows the well known exponential dependence

$$I_{off} \sim \exp\left[\frac{(V_g - V_{th})}{nV_t}\right]. \tag{18}$$

It increases by a factor of 10 for every 85 mV decrease in threshold voltage (i.e. about $2.3\times$ per generation). Other short-channel effects cause I_{off} to increase even more rapidly in scaled devices. This off-state leakage impacts the design of the feedback element of the preamplifier, which is sometimes required to have an equivalent resistance of hundreds of MΩ. To achieve such high equivalent resistance a much greater L/W ratio is needed. In addition, the use of switches with minimum L must be curtailed.

6.2. Noise from current sources

A secondary source of noise is the current source that supplies the input transistor. The noise contribution of the current source is proportional to the ratio $(g_{m2}/g_{m1})^2$ where g_{m1} and g_{m2} are the transconductance of the input device and current source respectively. To minimize this contribution, the ratio L/W of the current source transistor must be maximized while staying out of the linear region ($V_{DS} > V_{DS-sat}$); however, in order to preserve the dynamic range, the ratio must decrease as the supply voltage decreases. At the 0.13 μm generation, it will become impossible to design a current source that contributes less than 10% to the overall CSA noise.

6.3. Cross-talk/coupling

Coupling of digital and high-swing analog nodes back into the sensitive amplifier inputs is already a problem for mixed-signal circuits incorporating CSAs. In future technology generations, there will be competing mechanisms affecting cross-talk and coupling. First, V_{DD} scaling will limit the capacitive coupling by reducing dV/dt transients on the aggressor nodes. At the same time the higher

substrate doping will raise the source/drain junction capacitance of the aggressor nodes and also increase the back transconductance g_{mb} of the sensitive circuits. Hence, coupling of devices to the substrate will increase. In the interconnect, the use of low-ε dielectrics will reduce internode coupling and the availability of many (5–10) metal layers will allow more effective shielding.

6.4. Passive components

The most advanced CMOS processes are optimized for digital circuits and few foundries offer high quality passive components. We presented a new CSA structure that allows the resistor R_f in Fig. 1 to be replaced by a MOSFET in the triode region; a scaled copy of the R_f MOSFET provides pole-zero and non-linearity cancellation and has been experimentally verified [22–25]. A further modification to the classic CSA structure is shown in Fig. 8. Here MOS transistors replace both the feedback resistor R_f and the feedback capacitor C_f. The device M1 with source and drain shorted uses the gate to channel capacitance. This nonlinear capacitance is compensated by the series element $-n$M1. The pole formed by M1-M2 is cancelled by the zero from the series transistors, and transient non-linearity is likewise cancelled. The overall circuit injects a charge $-nQ_{in}$ into the second amplifying stage. In Fig. 8 both amplifiers A1, A2 must have virtual ground inputs at the same DC potential. For the second stage feedback impedance Z_{f2}, a linearized MOS resistor can be used for current to voltage conversion, his noise contributing as $1/n^2$ to the input equivalent parallel noise. The resulting circuit can perform amplification and filtering of low charge signals without relying on on-chip high quality passive components.

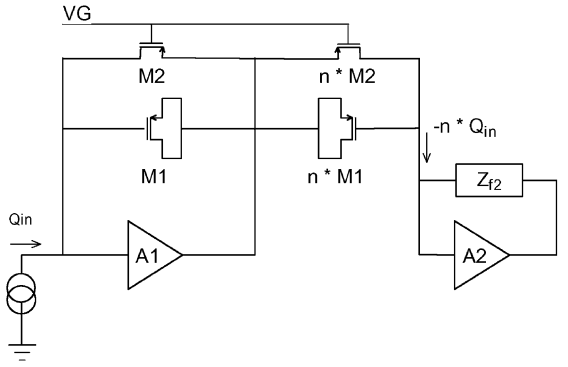

Fig. 8. Method of charge amplification using MOSFETs only.

7. Summary and outlook

CMOS charge sensitive amplifiers are now in widespread use in all methods of radiation detection. CSA design will be strongly influenced by the effects of technology scaling on device intrinsic noise sources, operating point, supply voltage, and available passive components.

In a survey of experimental noise measurements we have found that the series white and $1/f$ noise sources have remained well-behaved down to the 0.25 μm generation. Little evidence of hot-carrier induced excess noise is observed for devices biased at low V_{DS}, and PMOS retains its $1/f$ noise advantage over NMOS. Process modifications proposed for deep submicron devices, however, have the potential to drastically increase the $1/f$ noise. At the limits of CMOS scaling ($L_{min} < 0.1$ μm), gate tunneling current will become a significant source of parallel noise in MOSFETs.

In scaled devices, the onset of weak inversion occurs at ever-increasing current density. This shift of the operating region complicates the capacitive matching of the input FET to the detector. We derived a generalized solution of the optimum dimensioning problem and illustrated its effects on several realistic detector front-end designs.

Total ENC is expected to improve slowly with scaling, barring unforeseen catastrophic changes in the intrinsic FET noise sources. The higher f_T and reduced V_{DD} in scaled technologies can be exploited to reduce the power dissipated in the front transistor. However, supply voltage scaling limits the achievable signal-to-noise ratio, since the maximum signal swing will be strongly reduced.

Other complications that will arise as CMOS downscales are off-state leakage of switches and restricted availability of high quality passive components. Nevertheless, the potential for system-level integration in CMOS is of enormous

benefit for many applications, ensuring the continued productive use of this technology for detecting and manipulating small quantities of charge.

Acknowledgements

The authors would like to acknowledge G. Anelli, A. Rivetti, M. Campbell, and W. Snoeys (CERN) for sharing prepublication data on 0.25 μm devices; A. Kandasamy, D. Zhao, and J. Triolo (BNL) for helping with the noise measurements in Sections 4 and 5. Radeka (BNL) for help and encouragement with this project.

Appendix A. Generalized MOSFET matching conditions

Let us consider how to minimize the white series component of the equivalent noise charge. We rewrite Eq. (2) as

$$\mathrm{ENC}_{\mathrm{sw}}^2 = \frac{\alpha(C_{\mathrm{det}} + C_{\mathrm{g}})^2}{g_{\mathrm{m}}(C_{\mathrm{g}})} \tag{A.1}$$

where $\alpha = a_1 2kT\gamma/t_{\mathrm{s}}$, and where we allow the transconductance g_{m} to be an arbitrary function of the MOSFET gate capacitance C_{g}. (In this appendix, the drain current I_{D} is assumed to be held constant). Now (A.1) is minimized when

$$\frac{d(\mathrm{ENC}_{\mathrm{sw}}^2)}{d(C_{\mathrm{g}})} = \frac{2\alpha(C_{\mathrm{det}} + C_{\mathrm{g}})g_{\mathrm{m}} - \alpha(C_{\mathrm{det}} + C_{\mathrm{g}})^2 g'_{\mathrm{m}}}{g_{\mathrm{m}}^2} = 0. \tag{A.2}$$

Solving (A.2) for the optimum C_{g}, we obtain

$$2g_{\mathrm{m}} = (C_{\mathrm{det}} + C_{\mathrm{g,opt}})g'_{\mathrm{m}}. \tag{A.3}$$

Let us now consider the three regions of g_{m} behavior, as discussed in Section 5.1. First, in the strong inversion region we use Eq. (3b)

$$g_{\mathrm{m}} = \sqrt{2\mu C_{\mathrm{ox}} \frac{W}{L} I_{\mathrm{D}}} = \sqrt{\frac{2\mu I_{\mathrm{D}}}{L^2}}\sqrt{C_{\mathrm{g}}}. \tag{A.4}$$

Substituting (A.4) into (A.3) gives

$$2\sqrt{2\mu I_{\mathrm{D}}/L^2}\sqrt{C_{\mathrm{g,opt}}} = \frac{(C_{\mathrm{det}} + C_{\mathrm{g,opt}})\sqrt{2\mu I_{\mathrm{D}}/L^2}}{2\sqrt{C_{\mathrm{g,opt}}}}$$

$$4C_{\mathrm{g,opt}} = C_{\mathrm{det}} + C_{\mathrm{g,opt}}$$

$$C_{\mathrm{g,opt}} = C_{\mathrm{det}}/3 \tag{A.5}$$

which is the well known matching condition cited in Refs. [18–20].

In the velocity saturated case, Eq. (3c) applies

$$g_{\mathrm{m}} = \frac{C_{\mathrm{g}} v_{\mathrm{sat}}}{L}. \tag{A.6}$$

Solving (A.3) using (A.6) gives

$$C_{\mathrm{g,opt}} = C_{\mathrm{det}} \tag{A.7}$$

as the matching condition for velocity saturation.

In weak inversion, transconductance $g_{\mathrm{m}} = I_{\mathrm{D}}/(nV_{\mathrm{t}})$ is independent of C_{g}. In this case, Eq. (A.1) suggests that the minimum noise will occur when $C_{\mathrm{g}} = 0$. However, as the channel width is decreased towards zero, the resulting increase in current density eventually brings the device into strong inversion where $g_{\mathrm{m}} = I_{\mathrm{D}}/(nV_{\mathrm{t}})$ no longer applies. At that point, any further decrease in C_{g} will cause the noise to increase. Therefore, the optimum width will be the one for which the device is just at the boundary between weak and strong inversion.

Let us now derive the limits under which strong inversion matching is appropriate. In strong inversion the one-third-rule (A.5) applies, so we can write

$$C_{\mathrm{g}} = C_{\mathrm{ox}} WL = C_{\mathrm{det}}/3$$

$$W = C_{\mathrm{det}}/3C_{\mathrm{ox}}L. \tag{A.8}$$

Now the drain current can be written using the standard square-law formula

$$I_{\mathrm{D}} = \frac{\mu C_{\mathrm{ox}}}{2}\frac{W}{L}(V_{\mathrm{gs}} - V_{\mathrm{T}})^2$$

$$= \frac{\mu C_{\mathrm{det}}}{6L^2}(V_{\mathrm{gs}} - V_{\mathrm{T}})^2 \tag{A.9}$$

where V_{T} is the transistor's threshold voltage. From (A.9) we can determine the operating point

of the gate voltage

$$(V_{gs} - V_T) = \sqrt{\frac{6I_D L^2}{\mu C_{\rm det}}}. \quad (A.10)$$

In strong inversion operation $(V_{\rm gs} - V_{\rm T})$ must be within certain limits

$$\frac{v_{\rm sat} L}{\mu} > (V_{\rm gs} - V_{\rm T}) > 2nV_{\rm t} \quad (A.11)$$

where $V_{\rm t} = kT/q$ is the thermal voltage. If the left-side inequality is violated, the device is in velocity saturation. The right-hand inequality defines the weak inversion–strong inversion boundary [21].

Combining (A.11) with (A.10), we find after some algebra

$$\frac{6\mu}{v_{\rm sat}^2} < \frac{C_{\rm det}}{I_{\rm D}} < \frac{3L^2}{2\mu(nV_{\rm t})^2}. \quad (A.12)$$

Thus, we have established that the proper matching condition is determined solely by the ratio of $C_{\rm det}$ to $I_{\rm D}$. When this ratio is large, the input transistor dimensions must be set to put the device at the boundary of weak inversion. When it is small the MOSFET will be velocity saturated and should have $C_{\rm g} = C_{\rm det}$. Intermediate values call for the strong inversion condition $C_{\rm g} = 1/3\, C_{\rm det}$. Of course, in real devices the transitions between the regions of operation will be gradual.

Eq. (A.12) illustrates the fact that strong inversion bias points will become less common as successive generations of CMOS scale towards smaller channel length. As an example, let us compare the strong inversion limits for *n*-channel devices in 2 and 0.18 μm technologies. For 2 μm NMOS, Eq. (A.12) gives

$$3.7\times10^{-11} < C_{\rm det}/I_{\rm D} < 1.1\times10^{-7}\, F/A.$$

Thus, at a typical drain current $I_{\rm D}$ of 250 μA the strong inversion matching condition applies for detector capacitances $C_{\rm det}$

$$9.3\ {\rm fF} < C_{\rm det} < 26.3\ {\rm pF}.$$

For 0.18 μm technology, on the other hand, (A.12) tells us that the strong inversion rule is valid only over the detector capacitance range

$$9.3\ {\rm fF} < C_{\rm det} < 210\ {\rm fF}.$$

Appendix B. ENC scaling

Starting from Eq. (2) we can write

$${\rm ENC}_{\rm sw}^2 = \frac{a_1 2kT\gamma}{g_{\rm m} t_{\rm s}} (C_{\rm det} + C_{\rm g})^2. \quad (B.1)$$

In strong inversion Eq. (B.1) is minimized for $C_{\rm g} = C_{\rm det}/3$

$${\rm ENC}_{\rm sw}^2 = \frac{a_1 2kT\gamma}{g_{\rm m} t_{\rm s}} \left(\frac{4}{3} C_{\rm det}\right)^2. \quad (B.2)$$

We now substitute $C_{\rm ox} WL = C_{\rm g}/L^2 = C_{\rm det}/3L^2$. into expression (3b) for strong inversion transconductance

$$g_{\rm m} = \sqrt{2\mu C_{\rm det} I_{\rm D}/3L^2}. \quad (B.3)$$

Combining (B.2) and (B.3) gives

$$\begin{aligned} {\rm ENC}_{\rm sw}^2 &= \frac{a_1 2kT\gamma}{t_{\rm s}} \frac{\left(\frac{4}{3} C_{\rm det}\right)^2 \sqrt{3L^2}}{\sqrt{2\mu C_{\rm det} I_{\rm D}}} \\ &= \sqrt{\frac{3072}{162}}\, a_1 \gamma \frac{kTC_{\rm det}^{3/2} L}{t_{\rm s} \mu I_{\rm D}}. \end{aligned} \quad (B.4)$$

Combining the purely numerical terms and substituting $I_{\rm D} = P/V_{\rm DD}$ we arrive at

$${\rm ENC}_{\rm sw}^2 = \xi kTC_{\rm det} \frac{L_{\rm min}}{t_{\rm s}} \sqrt{\frac{C_{\rm det} V_{\rm DD}}{\mu P}}. \quad (B.5)$$

References

[1] E.A. Vittoz, IEEE J. Solid-State Circuits 20 (3) (1985) 657.
[2] M.J.M. Pelgrom, M. Vertregt, Nucl. Instr. and Meth. A 395 (1997) 298.
[3] V. Radeka, Ann. Rev. Nucl. Part. Sci. 38 (1988) 217.
[4] E. Gatti, P.F. Manfredi, Il Nuovo Cimento 9 (1986) 1.
[5] Semiconductor Industry Association, The National Technology Roadmap for Semiconductors, San Jose, CA, SIA, 1994; 1997.
[6] P.R. Gray, R.G. Meyer, Analysis and Design of Analog Integrated Circuits, Wiley, New York, 1984.
[7] A. van der Ziel, Proc. IRE 50 (1962) 1808.
[8] A. Birbas, D.P. Triantis, S.E. Plevridis, E.F. Tsakaset, IEEE Trans. Electron Devices 46 (6) (1999) 1253.
[9] P. Klein, IEEE Electron Device Lett. 20 (8) (1999) 399.

[10] R.P. Jindal, IEEE Trans. Electron Devices 33 (9) (1986) 1395.
[11] A. Abidi, IEEE Trans. Electron Devices 33 (11) (1986) 1801.
[12] G. Anelli, M. Campbell, M. Delmastro, F. Faccio, S. Florian, A. Giraldo, E. Heijne, P. Jarron, K. Kloukinas, A. Marchioro, P. Moreira, W. Snoeys, IEEE Trans. Nucl. Sci. 46 (6) (1999) 156.
[13] A. Rivetti, G. Anelli, F. Anghinolfi, M. Campbell, M. Delmastro, E. Heijne, F. Faccio, S. Florian, P. Jarron, K. Kloukinas, A. Marchioro, Analog design in deep sub-micron CMOS processes for LHC, Proceedings 1999 LEB, Snowmass, CO 25–28 (1999).
[14] J.C. Santiard, F. Faccio, Nucl. Instr. and Meth. A 380 (1996) 350.
[15] E. Simoen, C. Claeys, Solid State Electron. 43 (5) (1999) 865.
[16] E. Sheehan, P.K. Hurley, A. Mathewson, Microelectron. Reliab. 38 (6–8) (1998) 931.
[17] H.S. Wong, IEEE Trans. Electron Devices 43 (12) (1996) 2131.
[18] Z.Y. Chang, W. Sansen, Low-Noise Wide-Band Amplifiers in Bipolar and CMOS Technologies, Kluwer Academic Publishers, Dordrecht, 1991.
[19] W. Sansen, Z.Y. Chang, IEEE Trans. Circuits Syst. II 37 (11) (1990) 1375.
[20] G. Bertuccio, Micropower and Low-Noise Electronics for X- and γ-Ray Semiconductor Detectors, presented at the 11th International Workshop on Room Temperature Semiconductor X- and γ-Ray Detectors and Associated Electronics, Vienna, Austria, 11–15 (1999).
[21] K.R. Laker, W. Sansen, Design of Analog Integrated Circuits and Systems, McGraw-Hill, NewYork, 1994, p. 31.
[22] P. O'Connor, G. Gramegna, P. Rehak, F. Corsi, C. Marzocca, IEEE Trans. Nucl. Sci. 44 (3) (1997) 318.
[23] G. Gramegna, P. O'Connor, P. Rehak, S. Hart, Nucl. Instr. Meth. A 390 (1997) 241.
[24] G. Gramegna, P. O'Connor, P. Rehak, S. Hart, IEEE Trans. Nucl. Sci. 44 (3) (1997) 385.
[25] G. De Geronimo, P. O'Connor, Nucl. Instr. and Meth. A 421 (1999) 322.